

**High heritability of speech and language impairments in 6-year-old twins
demonstrated using parent and teacher report**

Dorothy V. M. Bishop, Glynis Laws, Caroline Adams, and Courtenay Frazier

Norbury

Department of Experimental Psychology

University of Oxford

Corresponding author:

Prof. D.V. M. Bishop,

Department of Experimental Psychology,

Tinbergen Building,

South Parks Road,

Oxford,

OX1 3UD.

tel: 01865 271369

fax: 01865 281255

email: dorothy.bishop@psy.ox.ac.uk

RUNNING HEAD: Heritability of speech and language impairments

Abstract

Previous twin studies have demonstrated high heritability of specific language impairment (SLI) when the diagnosis is based on psychometric testing. The current study measured the effectiveness of parent and teacher ratings of communication skills in identifying heritable language impairment. The Children's Communication Checklist was completed by parents and teachers of 6 year-old twins recruited from a general population sample. 130 twin pairs (65 MZ) were selected because at least one twin had low language skills at 4 years of age; a further 66 pairs (37 MZ) were a low risk group with no indication of language difficulties at 4 years. Internal consistency, inter-rater reliability, and validity in identifying language impairment were assessed for all CCC scales. CCC scales, especially those assessing structural language skills, were highly effective in identifying cases of language impairment, but agreement between parent and teacher ratings was modest. Genetic analysis revealed negligible environmental influence and substantial genetic influence on most scales. A rater-specific effects model was fit to the data to assess how far parents and teachers assess a common genetic factor on the CCC. Ratings of parents and teachers were influenced to some extent by the same child characteristics, but rater-specific effects were also evident, especially on scales measuring pragmatic aspects of communication. This study shows that there are strong genetic influences on both structural and pragmatic language impairments in children, and these can be detected using a simple checklist completed by parents or teachers.

Keywords: genetics, specific language impairment, pragmatics, checklist, assessment

Introduction

Most children acquire their first words around 12 months of age, start to string words together by 2 years, and are able to talk in long and complex sentences by 4 years of age. However, a minority are late in starting to speak, and when they do begin to talk, vocabulary and syntactic structure resemble that of a much younger child. Although many late talkers grow out of their difficulties in the preschool years, some have persisting problems throughout childhood, despite normal development of nonverbal abilities. Where obvious causal factors, such as peripheral hearing loss or physical handicap, are ruled out, and the child does not have more general intellectual difficulties or a recognised medical syndrome, the diagnosis of specific language impairment (SLI) is made. For many years the cause of SLI remained a mystery, but in the 1990s a series of twin studies and family aggregation studies demonstrated that SLI was a strongly heritable condition (see Bishop, 2002, for review). In general, these studies have based the diagnosis of SLI on standardized language tests.

The current study considers how effective parent and teacher reports of communication skills are in identifying heritable language impairments. There are two motivations for investigating this issue. The first is a practical one: molecular genetic studies of SLI require large sample sizes, and it is considerably less time-consuming to collect information from a parent or teacher checklist than to conduct an individual language assessment with each child. If heritable language impairments could be identified by checklist, this would improve the efficiency of molecular genetic investigations. Second, some aspects of language impairment are difficult to assess using traditional clinical instruments. Checklist ratings might provide valuable phenotypic information that is not readily obtained by more conventional means. As Bishop (1998) noted, most language assessments measure

the child's ability to comprehend, produce or repeat words or sentences, but they do not tap possible difficulties in the use of language in naturalistic contexts. For some children, pragmatic aspects of language pose particular problems: for instance, they may produce language in a stereotyped fashion, talk incessantly, fail to take a conversational partner's viewpoint into account, or be over-literal in their interpretation of language. In a survey of pupils attending special classes for children with SLI, Conti-Ramsden, Crutchley and Botting (1997) demonstrated that such pragmatic language impairments (PLI) were rated by teachers as particularly serious, yet they were identified only by clinical impression, and were not picked up on standardized assessments.

The instrument used in the current study is the Children's Communication Checklist (CCC), which was developed by Bishop (1998) as a means of standardizing professionals' observations of communication impairments in children. The CCC contains nine scales, five of which (inappropriate initiation, coherence, stereotyped language, use of context, rapport) were designed to evaluate pragmatic aspects of communication that are not easily assessed by more traditional means. Two further scales, speech and syntax, assess mastery of structural aspects of speech and language, and a final two scales, social interaction and interests, assess nonlinguistic behaviours that have been linked to pragmatic impairments. Sample items are given in Table 1, marked + for those assessing communication strengths, and – for those assessing weaknesses. The full list of CCC items is given in Bishop (1998).

Various studies with the CCC have demonstrated its sensitivity in identifying distinctive profiles of difficulty in children with evidence of language impairments (Bishop, 1998), neurodevelopmental disorders including autism and ADHD (Bishop and Baird, 2001, Geurts et al., 2004), conduct disorder (Gilmour, Hill, Place and Skuse, 2004) and Williams syndrome and Down syndrome (Laws and Bishop, 2004).

Bishop and Baird (2001) and Geurts et al. (2004) also compared CCC data provided by parents and professionals (mostly teachers) on the same children. Although data from both sets of informants showed sensitivity of CCC data to the child's clinical status, the agreement between parent and professional ratings on the pragmatic scales was not high (correlations ranged from around .3 to .6). This is a common finding when ratings are used to evaluate children's behaviour, and it prompts the question of whether the poor agreement reflects genuine differences in what is being rated, or is just a consequence of systematic bias (the tendency by some raters to give high or low ratings) or random error. This is a question we shall return to in this paper.

More recently, a new version of the checklist, the CCC-2, has been developed with a format more suited for parents (Bishop, 2003). Although the initial goal of both CCC and CCC-2 was to provide a means for distinguishing subtypes within a population of language-impaired children, research with the CCC-2 showed that a total score based on structural as well pragmatic subscales was highly effective at discriminating children with communication difficulties from typically-developing children (Norbury, Nash, Bishop, and Baird, 2004).

The current study was initiated before the CCC-2 had been developed, so the original version of the CCC was used. It was given to parents and teachers of 6-year-old children who were participating in a twin study concerned with the heritability of children's communication impairments. This study had two main goals:

1. The first goal was to establish the reliability and validity of the CCC with 6-year-old children. Although children as young as 6 years were included in the study by Bishop and Baird (2001), they were in the minority, and the psychometric characteristics of the checklist have not been established with children this young. We computed internal consistency, test-retest reliability, and inter-rater reliability,

and also considered how effective the CCC was at distinguishing children at risk of language impairment from low risk children.

2. The second goal was to compare similarities between monozygotic (MZ) and dizygotic (DZ) twins to discover whether the CCC was sensitive to heritable aspects of language impairment. Twins who grow up together may resemble each other in communication skills by virtue of their shared environment, in which case we expect to see significant correlations between ratings of twin 1 and twin 2 in a pair.

However, if genetic differences between individuals account for differences in children's communication skills, we would expect MZ twins, who share all their segregating genes, to resemble one another more closely than DZ twins, who share on average only 50% of their segregating genes. If there are low correlations between two members of a twin pair, then this indicates that neither shared environment nor shared genes affect variation, which may instead be due to factors unique to the child, or unreliability of the measures.

Because we used both parent and teacher ratings to assess the phenotype, we were able to partition genetic influences into those that are common to both raters and those that are rater-specific. This made it possible to assess whether disagreement between parent and teacher ratings arises because each type of rater assesses different but meaningful aspects of the phenotype, or whether the disagreement merely reflects systematic rater bias or random error.

Methods

Participants

Children were a subset of cases from the Twins Early Development Study (TEDS), an population sample of twins born in England and Wales (Trouton, Spinath, and Plomin, 2002). For the main TEDS sample, live twins born in 1994-1996 were

identified via the UK Office for National Statistics, and their parents were contacted when the twins were around a year old. Parents completed assessments of their children's language and nonverbal abilities at 2, 3 and 4 years of age (Dale, Price, Bishop, and Plomin, 2003). On the basis of parental report, children were identified as at risk of language impairment ('LI risk') if they had a poor score on any one of three indices: (i) child described as not yet talking in full sentences; (ii) vocabulary in lowest 10%, as judged from parental checklist; (iii) parent answered 'yes' to the question 'Do you have any concerns about your child's speech and language?' and selected the option 'his/her language is developing slowly' when asked to specify the nature of the concern. Of 5426 same-sex twin pairs with 4-year parental report data, 547 (10.6%) met criteria for low language at 4 years in one or both twins. A subset of LI risk children were selected for in-depth study and compared with a low risk group of twins who were selected on the basis that neither met criteria for LI risk at 4 years.

We excluded cases where the language impairment was associated with sensorineural hearing loss, physical handicap, autism, or another syndrome affecting cognitive development. Children who failed a hearing screen when assessed (average hearing threshold for frequencies 500 to 400 Hz higher than 26 dB in the better ear) were also excluded, as well as families where English was not the only language spoken in the home. The participants were selected to be white in order to reduce the possible effects of ethnic stratification in future molecular genetic studies, and the sample was restricted to same-sex pairs. The sample for this study was selected so that pairs where one or both children met criteria for LI risk constituted around 2/3 of the sample, and there were equal numbers of MZ and DZ pairs; however, twin concordance for LI pairs was not taken into account when selecting twin pairs, as this would have biased heritability estimates. The subset of children selected for in-depth

study (see Table 2) did not differ from the remainder of the sample in terms of socio-economic status.

Assessments

Twin children were seen individually in a quiet room at home or school for an assessment lasting around 90 minutes, which included the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999), three subtests from the Clinical Evaluation of Language Fundamentals – Revised (Semel, Wiig, and Secord, 1987) and the Children’s Nonword Repetition Test (Gathercole, Willis, Baddeley, and Emslie, 1994). Raw scores on these assessments were transformed to age-scaled scores on the basis of published norms, and then re-scaled to mean 100 and SD 15.

Parents and teachers of participating twins were given a copy of the Children’s Communication Checklist (Bishop, 1998) to complete for each twin, and asked to complete it and mail it back to the research team.

Results

Children with CCC data

CCC ratings for both twins were obtained from at least one rater for 95% of the sample. There were 77 MZ twin pairs and 70 DZ twin pairs with CCC ratings for both children from both a parent and a teacher. A further 18 MZ and 18 DZ pairs had parent ratings only, and 3 MZ and 1 DZ pairs had ratings from a teacher only. For those with teacher ratings, 60% of MZ and 60% of DZ twins were rated by the same teacher.

In some cases, respondents had failed to answer every questionnaire item. Teachers were particularly likely to respond ‘unable to judge’ for items on the Social Relationships and Interests scales. Where responses for a case were available for less

than 85% of the items for a specific subscale, the case was omitted from analyses involving that subscale. However, provided other subscales achieved the 85% criterion, then the scores for that case were used in other analyses. The Pragmatic Composite (PC) was calculated only for cases where valid data were available for all five pragmatic language subscales.

In the CCC-2, a General Communication Composite (GCC), formed by summing all the communication scales, was found to be useful in screening for communication impairments in children. To see if a comparable scale would be effective at identifying language impairments in the current sample, a GCC was formed by summing scales A to G.

Characteristics of the sample

Table 3 shows mean ages and data on the psychometric tests for children categorised according to LI risk status. This categorisation was made on the basis of each individual child's 4-year-old status, regardless of the co-twin's status. The LI risk and low risk groups differed significantly at the .05 level on all psychometric measures. Results from the psychometric tests were used to categorise children as having general delay if Performance IQ (PIQ) was less than 85. SLI was identified if PIQ was 85 or above, and scores on two or more of the four language tests (three CELF-R subtests or nonword repetition) were below 85 (i.e. below 16th centile). At 6 years of age, 31.9% of the LI risk group were categorised as having SLI and 13.1% as having general delay. In the low risk group, 10.9% of children had SLI at 6 years and 3.5% had general delay. Overall this represents a substantial excess of children with impairments in the LI risk group, $\chi^2(2) = 44.4$, $p < .001$. The LI risk and low risk groups did not differ in proportion of MZ twins: 49% in low risk group and 54% in LI risk group, $\chi^2(1) = 1.26$, $p = .26$. There were no gender differences when children

were categorised according to LI risk: for low risk, males = 52.2% and for LI risk, males = 58.1%, $\chi^2(1) = 1.37$, $p = .24$. However, when children were categorised according to 6 year status, males were more frequent in the SLI group: unaffected, males = 52.3%; SLI, males = 69.9%, general delay, males = 40.6%, $\chi^2(2) = 10.89$, $p = .004$.

Reliability and validity of CCC

Internal consistency of subscales and pragmatic composite

Table 4 shows the reliability of subscales and the two composites for parent- and teacher-completed CCCs. This shows comparable levels of internal consistency for scales completed by parents and teachers. In most cases, scales demonstrated acceptable levels of reliability of .7 or greater. The Interests subscale (I) was an exception with Cronbach's α of less than .5 and so was not used in further analyses. The PC and GCC both produced a Cronbach's α around 0.9 for parent- and teacher-completed questionnaires.

Inter-rater reliability

For cases where CCCs were completed by a parent and a teacher, inter-rater reliabilities for each of the subscales and the composites were calculated using the Pearson correlation. These values are shown in Table 4, and indicate that correlations exceed .5 only for scales A (Speech), B (Syntax) and D (Coherence), and for the GCC. For the other scales, agreement between parent and teacher ratings is generally weak (though in all cases above chance). These values are similar to those reported by Bishop and Baird (2001).

Validity: comparison of CCC scores for LI risk and low risk children

One indicator of validity of the CCC is its ability to distinguish children who are identified as having language impairments. Table 5 shows the mean CCC scores for children in relation to their 4-year-old language status. The effect size, d , is shown so that one can compare CCC scales with the psychometric tests in Table 3 in terms of their ability to distinguish LI risk from low risk children. Perhaps the most striking point to emerge from Table 5 is that the Speech, Syntax and Coherence scales, as well as the GCC, are at least as effective at discriminating LI risk from low risk groups as the psychometric tests. This is the case even when ratings are made by teachers, who played no part in assessing LI risk at 4 years.

Figure 1 shows GCCs for ratings by parents and teachers in relation to the child's language status at 6 years. The mean difference between impaired (SLI or general delay) and unaffected children is slightly greater than 1 SD for both parent and teacher ratings, indicating that the GCC is effective in identifying children who meet psychometric criteria for language impairment.

Genetic analysis

Twin-cotwin correlations

A preliminary estimate of genetic and environmental influences on CCC ratings can be obtained by scrutinising the pattern of twin-cotwin correlations for MZ and DZ twin pairs. Because maximum likelihood estimation assumes approximate normal distribution of the data, CCC scores were reflected and log transformed prior to analysis, so that the means on all scales, including the PC and GCC, were close to 1.0, and skewness was less than 1.0 for all variables. Table 6 shows the intraclass correlations between twin 1 and twin 2 for MZ and DZ pairs when assessed by parents and teachers, both for children rated by the same person, and for children rated by different people.

An estimate of the genetic influences on the phenotype can be obtained by doubling the difference between MZ and DZ correlations for twins rated by the same person. It is evident that heritabilities of parental ratings are very high for the three scales that gave particularly strong differentiation between LI risk and low risk twins, A (Speech), B (Syntax) and D (Coherence). The GCC also shows strong genetic influence. The trend is similar when the same teacher rates both twins (columns headed teacher A – teacher A). The question arises as to whether heritability estimates could be distorted if raters have difficulty differentiating between MZ twins, who are very similar in appearance, and so tend to give them the same ratings. However, this could not explain why there should be large MZ-DZ differences in correlations on some scales for pairs where the twin and co-twin were rated by different teachers (columns headed teacher A – teacher B). In addition, there is little evidence in these data for systematic bias arising because some raters tend to rate both twins 'good' and others to rate both twins 'poor'. This would elevate correlations for both MZ and DZ twins, leading to overestimation of environmental influences that affect both twins; for most measures in Table 6, it would appear that such environmental influences (estimated by twice the DZ correlation minus the MZ correlation) are negligible. The final pair of columns in Table 6 gives the intraclass correlation for twin 1 rated by a parent and twin 2 rated by a teacher (where twin 1 and twin 2 are assigned at random). It should not surprise us that these correlations are relatively low, as they will be constrained by the inter-rater reliability between parents and teachers (see Table 4); i.e., the correlation between a parent rating twin 1 and a teacher rating twin 2 is unlikely to be higher than the correlation between these raters when rating the same child. The correlations for DZ twins are all close to zero, whereas those for MZ twins tend to be higher. Overall, the pattern of correlations

suggests that parents and teachers are both rating heritable traits, and that, though they are not rating exactly the same traits, there is some overlap.

Model-fitting

The model shown in Figure 2 was fit to the raw data to quantify these impressions. This is equivalent to a model developed by Hewitt, Silberg, Neale, Eaves, and Erickson (1992), which they termed the ‘psychometric model’, but we have termed it the ‘rater-specific effects’ model, to avoid any confusion with the psychometric tests in our study. In this model, observed ratings by parent and teacher are determined both by common characteristics of the child that contribute to both sets of ratings, and by rater-specific observations. The lower half of model shows the ratings for parents and teachers (shown in oblongs in Figure 2) linked in terms of the standard ACE model, in which covariances between twins rated by the same person are explained in terms of additive genetic, common environment and nonshared environmental influences, termed, A_p , C_p , E_p for parent ratings, and A_t , C_t , and E_t for teacher ratings respectively. If each set of ratings were entirely independent, the lower half of the model would suffice to describe the data. However, it is anticipated that parents and teachers are to some extent rating the same phenotype. The common phenotype is depicted in the ovals labeled “Phenotype twin 1” and “Phenotype twin 2”, and is influenced by genetic, shared environmental and unique environmental influences, labeled A , C and E , respectively. The proportion of variance in observed ratings that is explained by the common phenotype is denoted by z_p . When z_p is zero, the model becomes equivalent to two separate ACE models, one specific for parent ratings and one for teacher ratings.

There are two potential sources of influence on the rater-specific terms in the model: situation effects, and rater effects. To illustrate this point with a fictional

example, suppose that a child has a tendency to ignore conversational overtures when in the company of a group of other children, but not when in a small family group: this would show up as a rater-specific effect because it is an aspect of the phenotype that is more likely to be evident in the school situation than at home. This behaviour could potentially be determined predominantly by genetic factors (showing up in the A_i term) or by shared or nonshared environment (C_t or E_t). However, situation-specific effects are not the only source of influence on rater-specific terms. Another important potential factor that has been discussed in the literature is rater bias. Any systematic bias in the tendency of individual raters to give high or low ratings to a twin pair will inflate the twin-cotwin correlations and so show up as an effect of common environment. Because raters are independent of one another, such an effect will influence the C_p or C_t terms but not the C term for the common phenotype.

Although labelled as genetic terms, the A_p and A_i could also be affected by rater effects. Spinath and Angleitner (1998) noted that contrast effects can mimic dominant genetic effects if raters are biased to rate MZ twins as similar to one another, and DZ twins as different from one another. This kind of zygosity-dependent contrast effect may be suspected when, as in parent ratings on scales A, B and D, the intraclass correlations for DZ pairs are substantially lower than those for MZ twins, and are close to zero. However, such effects would be specific to a rater, and so would influence only the rater-specific genetic terms.

One way of distinguishing between situation and rater effects on the A_p , A_i , C_p and C_i terms is to consider patterns of covariance for children rated by the same or different raters. Rater bias or rater contrast effects cannot affect twin-cotwin similarity for children rated by different teachers, whereas situation effects can. Thus the probable source of "rater-specific" effects can be investigated by testing whether

paths in the lower half of Figure 2 are comparable for children rated by the same or different teachers.

Van der Valk et al. (2001) noted that, in this model, the A, C and E factors loading on the common phenotype contain only reliable trait variance, so the E term does not include error of measurement, but only genuine idiosyncratic environmental effects. Measurement error will be subsumed in the E_p and E_t terms.

For the current data, modelling was based on a script from Posthuma et al. (2000), which is run using the program Mx (Neale, Boker, Xie, and Maes, 1999). This was modified to allow for the fact that the dataset was subdivided into two groups: those who were rated by the same teacher, and those rated by different teachers. The sample size was too small to give meaningful estimates of sex effects, and so boys and girls were treated together as a single group.

A correction for ascertainment bias was included in the model, to adjust covariances to take into account the fact that the ratio of language risk to low risk twin pairs in the sample was 2:1, whereas it was 1:9 in the population from which the twins were drawn. After normalising the data with reference to the estimated true population mean and SD, we followed the method described by Wade, Neale, Lake, and Martin (1999), using the Weight term in Mx to adjust covariances on the assumption that both groups were drawn from a population in which the true mean was zero, and that pairs with a language-risk twin represented 10% of the population, and those with no language-risk cases represented 90%. In practice, this is achieved by specifying all estimated means to be zero, and by including in each group of twins the command:

Weight (I-\mnor(C_M_T_T_Q))~ ;

in the Mx script, where I is one, C is the estimated covariance matrix, M is a 1 x 4 matrix of means (all zero), T is a 1 x 4 matrix of thresholds, expressed as a z-score

that selects the top 10% of the population ($= 1.28$) and Q is a 1×4 matrix of ones (when selecting language risk pairs, where one or both twins score above threshold), or zeroes (when selecting low risk pairs, where neither scores above threshold).

Standardized parameter estimates for each scale are shown in Table 7, for a version of the model where paths were equated for twins rated by the same or different teachers, except for the a_t , c_t and e_t paths, which were free to vary. Rater-specific effects are shown for twin pairs rated by the same teacher. As discussed further below, rater-specific effects a_t and c_t were mostly low and non-significant for twins rated by different teachers.

To illustrate how model estimates are interpreted, consider the total genetic influence on parental ratings for a scale (see Table 7, row labelled ‘total genetic’). This corresponds to the expected size of genetic effect on parental ratings if entered into a standard ACE model, and is computed as $(z_p.a) + a_p$. The analogous total shared environmental influence (not shown) would be $(z_p.c) + c_p$, and the total non-shared environmental influence would be $(z_p.e) + e_p$. The estimates have been squared and standardized, so the sum of all three influences is 1.0. The parameter z_p estimates the proportion of variance in a rating that is due to the common phenotype.

The three scales that gave especially good differentiation between LI risk and low risk children (A, B and D) show a similar pattern: There is a relatively strong contribution from the common phenotype, and the common phenotype is under substantial genetic influence. The other scales, especially scale E, have less contribution from the common phenotype and so are more subject to rater-specific influences. In general the pattern of findings is similar for parent and teacher ratings.

Table 7 also shows the p-values obtained when the model was modified by dropping all environmental terms except rater-specific non-shared environment terms (which are needed to represent measurement error). All p-values are substantially

greater than .05, indicating that these terms can be dropped from the model without impairing the fit. Thus environmental influences shared by the twins, and unique environmental influences affecting one twin (other than measurement error) have negligible effects on ratings. Note that the rater-specific environment terms will include effects of systematic rater bias, i.e., the tendency of a given rater to inflate or minimize a child's communicative difficulties. The fact that the environmental terms are non-significant indicates that any rater-specific effects on results cannot be explained by such consistent biases.

The model in Figure 2 was next modified by setting all terms relating to the common phenotype (a, c and e) to zero, to test the significance of the common phenotype. This dramatically impaired the fit of the model for all measures, indicating that a model incorporating a common phenotype is necessary to account for the pattern of covariance in the data.

A further analysis looked at the effect of dropping the specific genetic terms from the model. For most of the scales, this significantly impaired the model fit. As noted above, interpretation of the specific genetic term is not straightforward: it could reflect genuine genetic influences on a phenotype that is specific to a given setting, but it will also be affected by zygosity-specific rater contrast effects. However, the latter will only occur for teacher ratings when two twins are rated by the same teacher. Therefore, we can argue that if the a_i term is nonsignificant for twins rated by different teachers, then this suggests it largely reflects rater contrast effects. The model was accordingly re-run with a_i and c_i terms fixed to zero for twins rated by different teachers. The fit was not significantly affected except for scale E (stereotyped conversation) and the pragmatic composite. In a further analysis (not shown in Table 7), we equated estimates of a_i and c_i terms for twin pairs assessed by same or different teachers and found this significantly impaired the fit of the model

for all scales except scale A (where rater-specific effects are small in magnitude). Thus for most scales, we can conclude that the significant a_i term is indicative of zygosity-specific contrast effects, rather than genetic influences on behavior that is seen only in the school setting.

The sample size in this study was relatively small, raising the question of whether the failure to obtain significant environmental terms might be due to low power. Power was assessed on the basis of a simulation in which the true value of the common a^2 , c^2 and e^2 terms were set to .8, .1 and .1, and all rater-specific terms were set to .2; this simulation was tested against a model in which all environmental terms were fixed to zero. A sample of 160 twin pairs has a power of approximately .8 to detect such effects, suggesting that the failure to find any environmental effects in the current study is unlikely to be due to the small sample size.

Discussion

The CCC was developed mainly to assess *pragmatic* aspects of language, which are hard to evaluate using conventional tests, yet it was the scales assessing *structural* aspects of language that gave the most striking results in this analysis. Three scales stood out: A (speech), B (syntax) and D (coherence). They showed substantial effect sizes when comparing LI risk and low risk children, they had reasonable inter-rater reliability, and gave high estimates of heritability.

The coherence scale was originally intended to identify children who might have poor discourse skills even if structural skills were adequate, but in a previous study it was found that children with classic SLI obtained much lower rating than typically-developing children (Bishop and Baird, 2001). This is perhaps not surprising: children who have limited ability to express themselves using complex syntax are likely to be rated as below average in their ability to construct coherent

discourse. In the CCC-2 the coherence scale was recategorised so that it was no longer treated as a measure of pragmatic competence, but was grouped with the other structural scales (speech and syntax).

It should perhaps be expected that the scales assessing pragmatic aspects of communication gave weaker inter-rater reliability, because these assess behaviours that are more situation-dependent. Inter-rater agreement was particularly weak for scales measuring inappropriate initiation (C) and stereotyped conversation (E). Nevertheless, the rater-specific effects model indicated that these scales are sensitive to heritable aspects of communication. The significant rater-specific effects suggest that different raters may rate different facets of the child, but it should be noted that a tendency for raters to emphasise similarities between MZ twins, and differences between DZ twins will also contribute to rater-specific effects.

Overall, these results are encouraging in suggesting that ratings by parents and teachers can provide important information to help identify heritable aspects of language impairment. The drawback of ratings is that they are far more subjective than psychometric assessments, but they have the advantage that they enable one to obtain impressions from a person who knows the child well, and allow one to evaluate behaviours that may be difficult to elicit in a clinical setting. It was noteworthy that some of the CCC scales not only gave substantial genetic effects, but also did at least as well as conventional language tests in discriminating between children with and without language risk. This suggests that the CCC might pick up qualitative aspects of communication that are key features of language impairment. Language tests, although more objective than ratings, may be more influenced by factors other than the child's genetic predisposition to language impairment: e.g. home language environment, or attention or motivation.

The results obtained here suggest that parent and teacher ratings can be used with confidence as a measure of the phenotype in molecular genetic studies. The CCC has now been superseded by CCC-2, which covers the same broad content as the CCC, but uses a different response format, with wording more suited to parent informants. It is hoped that CCC-2 will prove to be as effective in identifying heritable language impairments as its predecessor, and we are currently gathering CCC-2 data with the same twin sample at 9 years of age to provide data on this point.

Acknowledgements

We thank the twins and their families and teachers who participated in this research. This study would not have been possible without generous assistance of Robert Plomin, Bonamy Oliver, Alexandra Trouton and other staff from the Twins Early Development Study. Thanks are also due to Barbara Arfe and Lesley Bretherton for assistance with data collection, and to Michael Neale for advice on methods of correcting for ascertainment bias. This research was supported by a programme grant from the Wellcome Trust.

References

- Bishop, D. V. M. (1998). Development of the children's communication checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *J. Child Psychol. Psychiat.* **39**: 879-891.
- Bishop, D. V. M. (2002). The role of genes in the etiology of specific language impairment. *J. Commun. Dis* **35**: 311-328.
- Bishop, D. V. M., and Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication: use of the Children's Communication Checklist in a clinical setting. *Dev. Med. Child Neurol.* **43**: 809-818.
- Conti-Ramsden, G., Crutchley, A., and Botting, N. (1997). The extent to which psychometric tests differentiate subgroups of children with SLI. *J. Speech Lang. Hear. Res.* **40**: 765-777.
- Dale, P. S., Price, T. S., Bishop, D. V. M., and Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient delay at 3 and 4 years. *J. Speech Lang Hear. Res.* **46**: 544-560.
- DeFries, J. C., and Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behav. Genet.* **15**: 467-473.
- Gathercole, S. E., Willis, C., Baddeley, A. D., and Emslie, H. (1994). The children's test of nonword repetition: a test of phonological working memory. *Memory* **2**: 103-127.
- Gilmour, J., Hill, B., Place, M., and Skuse, D. H. (2004). Social communication deficits in conduct disorder: a clinical and community survey. *J. Child Psychol. Psychiatry* **45**: 967-978.
- Geurts, H. M., Verte, S., Oosterlaan, J., Roeyers, H., Hartman, C. A., Mulder, E. J., Berckelaer-Onnes, I. A., and Sergeant, J. A. (2004). Can the Children's

- Communication Checklist differentiate between children with autism, children with ADHD, and normal controls? *J. Child Psychol. Psychiatry* **45**: 1437-1453.
- Hewitt, J. K., Silberg, J. L., Neale, M. C., Eaves, L. J., and Erickson, M. (1992). The analysis of parental ratings of children's behavior using LISREL. *Behav. Genet.* **22**: 293-317.
- Laws, G., and Bishop, D. V. M. (2004). Pragmatic language impairment and social deficits in Williams syndrome: a comparison with Down's syndrome and specific language impairment. *Int. J. Lang. Commun. Disord.* **39**: 45-64.
- Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (1999). *Mx: Statistical Modeling, 5th Edition*. VCU Box 900126, Richmond, VA 23298: Department of Psychiatry.
- Norbury, C. F., Nash, M., Bishop, D. V. M., and Baird, G. (2004). Using parental checklists to identify diagnostic groups in children with communication impairment: A validation of the Children's Communication Checklist - 2. *Int. J. Lang. Commun. Disord.* **39**: 345-364.
- Posthuma, D., Beem, A., de, G., EJ., van, B., GC., von, H., JB., Iachine, I., and Boomsma, D. (2003). Theory and practice in quantitative genetics. *Twin Res.* **6**: 361-376.
- Rust, J., Golombok, S., and Trickey, G. (1993). *Wechsler Objective Reading Dimensions*. Sidcup, UK: Psychological Corporation.
- Semel, E. M., Wiig, E. H., and Secord, W. (1987). *Clinical Evaluation of Language Fundamentals - Revised*. San Antonio, Texas: Psychological Corporation.
- Spinath, F. M., & Angleitner, A. (1998). Contrast effects in Buss and Plomin's EAS questionnaire: a behavioral-genetic study on early developing personality traits assessed through parental ratings. *Pers. Ind. Diff.* **25**: 947-963.

- Trouton, A., Spinath, F. M., and Plomin, R. (2002). Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood. *Twin Res.* **5**: 444-448.
- Van der Valk, J. C., Van den Oord, E. J. C. G., Verhulst, F. C., and Boomsma, D. I. (2001). Using parental ratings to study the etiology of 3-year-old twins' problem behaviors: different views or rater bias? *J. Child Psychol. Psychiatry* **42**: 921-931.
- Wade, T., Neale, M. C., Lake, R. I. E., and Martin, N. G. (1999). A genetic analysis of the eating and attitudes associated with bulimia nervosa: dealing with the problem of ascertainment in twin studies. *Behav. Genet.* **29**: 1-10.
- Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio: Psychological Corporation.

Table 1

Sample items from CCC. Items marked + assess communicative strengths, and those marked – assess abnormalities or deficits

scale	-/+	sample item
A. Speech	-	production of speech sounds seems immature, like that of a younger child, e.g. says things like: "tat" for "cat", or "chimbley" for "chimney", or "bokkle" for "bottle"
B. Syntax	+	can produce long and complicated sentences such as: "When we went to the park I had a go on the swings"; "I saw this man standing on the corner"
C. Inappropriate Initiation	-	talks to anyone and everyone
D. Coherence	+	can give an easy-to-follow account of a past event such as a birthday party or holiday
E. Stereotyped conversation	-	often turns the conversation to a favourite theme, rather than following what the other person wants to talk about
F. Use of context	+	can understand sarcasm (e.g., will be amused rather than confused when someone says "isn't it a lovely day!" when it is pouring with rain).
G. Rapport	-	ignores conversational overtures from others (e.g. if asked "what are you making?" just continues working as if nothing had happened)
H. Social Relationships	+	is popular with other children
I. Interests	-	has a large store of factual information: e.g., may know the names of all the capitals of the world, or the names of many varieties of dinosaurs.

Table 2

Numbers of twin pairs selected for in-depth study in relation to zygosity, gender and

LI risk status

	twins with LI risk			
	neither twin	one twin	both twins	total
MZ female	17	8	16	41
DZ female	16	22	9	47
MZ male	19	17	24	60
DZ male	13	23	12	48
total	65	70	61	196

Table 3

Mean (SD) age and psychometric test scores for low risk and LI risk children

	low risk		LI risk		effect size (d)
	N*	mean (SD)	N	mean (SD)	
age (yr)	201	6.5 (0.18)	191	6.5 (0.19)	-
PIQ	201	100.9 (11.15)	191	97.09 (10.52)	0.34
VIQ	201	100.7 (13.02)	191	91.9 (12.89)	0.65
CELF-R Listening to paragraphs	201	99.9 (13.40)	191	93.7 (15.89)	0.41
CELF-R Sentence structure	201	99.2 (13.00)	189	92.2 (12.67)	0.53
CELF-R Recalling sentences	200	96.5 (12.35)	185	86.0 (13.85)	0.75
Nonword repetition	198	96.0 (17.23)	175	84.8 (18.93)	0.59

*Where N is less than full sample size, child refused test or was too unintelligible to score

Table 4

Internal consistency (Cronbach's α) and inter-rater reliability (Pearson r) for parent and teacher ratings

CCC scale	N items	Parent rating		Teacher rating		Parent-teacher agreement	
		α	N cases	α	N cases	Pearson r	N cases
A. Speech	11	.83	351	.86	285	.61	294
B. Syntax	4	.74	360	.77	298	.52	284
C. Inappropriate initiation	6	.75	356	.70	280	.25	266
D. Coherence	8	.86	332	.90	271	.56	287
E. Stereotyped conversation	8	.74	338	.78	269	.17	273
F. Use of context	8	.65	333	.66	249	.39	289
G. Rapport	8	.73	353	.78	286	.30	289
H. Social relationships	10	.63	334	.70	277	.36	285
I. Interests	7	.41	342	.45	177	.32	254
PC	38	.89	284	.91	200	.42	233
GCC	53	.91	272	.93	194	.53	226

Table 5

Mean (SD) for low risk and LI risk children on CCC, with effect size (d) shown for significant differences

	low risk		LI risk		
	N	mean (SD)	N	mean (SD)	d
Parent rating					
A. Speech	185	34.2 (2.85)	181	30.3 (4.77)	0.88
B. Syntax	183	31.4 (1.10)	175	30.3 (1.92)	0.69
C. Inappropriate initiation	182	26.2 (2.66)	174	26.0 (2.74)	-
D. Coherence	183	33.8 (2.23)	179	31.0 (3.95)	0.82
E. Stereotyped conversation	180	26.0 (2.89)	176	25.7 (3.12)	-
F. Use of context	180	28.6 (2.48)	175	27.0 (2.86)	0.56
G. Rapport	185	32.1 (2.06)	179	30.5 (2.99)	0.62
H. Social Relationships	183	32.0 (2.04)	172	31.6 (2.55)	-
PC	175	146.8 (9.00)	167	140.3 (11.49)	0.60
GCC	173	212.6 (10.43)	166	201.0 (16.10)	0.79
Teacher rating					
A. Speech	162	34.1 (3.13)	140	29.9 (5.08)	0.91
B. Syntax	162	31.3 (1.32)	138	30.0 (2.24)	0.70
C. Inappropriate initiation	148	27.4 (2.08)	134	27.5 (2.35)	-
D. Coherence	161	33.5 (3.23)	137	30.4 (4.58)	0.73
E. Stereotyped conversation	153	28.0 (2.55)	136	27.3 (3.09)	-
F. Use of context	155	29.3 (2.23)	137	28.1 (2.77)	0.49
G. Rapport	161	31.4 (2.69)	138	30.1 (3.26)	0.43
H. Social Relationships	162	32.0 (2.40)	141	31.4 (2.78)	-
PC	137	150.1 (9.08)	122	143.2 (12.18)	0.62
GCC	135	216.1 (11.00)	119	203.4 (17.13)	0.82

Table 6

Intraclass correlations for twin 1 vs. twin 2 (N pairs) in relation to zygosity and rater

scale	parent		teacher A – teacher A		teacher A – teacher B		parent - teacher	
	MZ	DZ	MZ	DZ	MZ	DZ	MZ	DZ
A. Speech	.69 (94)	.03 (89)	.82 (47)	.08 (41)	.72 (28)	.28 (29)	.38 (75)	.13 (72)
B. Syntax	.79 (92)	.15 (84)	.68 (45)	.43 (40)	.76 (30)	.28 (28)	.43 (72)	0 (70)
C. Inappropriate init.	.76 (92)	.32 (83)	.52 (43)	.30 (40)	.48 (25)	.18 (23)	.06 (68)	-.06 (65)
D. Coherence	.77 (90)	.19 (89)	.90 (45)	.60 (40)	.48 (28)	.10 (28)	.39 (70)	.06 (70)
E. Stereotyped	.77 (91)	.47 (84)	.85 (46)	.42 (38)	.52 (27)	.52 (22)	.16 (72)	-.09 (63)
F. Context	.71 (89)	.40 (86)	.89 (46)	.62 (38)	.57 (27)	.03 (24)	.35 (72)	-.12 (66)
G. Rapport	.58 (93)	.16 (87)	.82 (47)	.49 (40)	.40 (30)	-.13 (27)	.26 (75)	-.07 (69)
H. Social	.47 (90)	.22 (84)	.89 (47)	.60 (41)	.46 (31)	.43 (27)	.18 (74)	.04 (68)
PC	.77 (85)	.41 (79)	.93 (40)	.69 (36)	.83 (18)	.14 (17)	.30 (60)	-.13 (56)
GCC	.77 (85)	.34 (76)	.92 (40)	.56 (34)	.78 (17)	.17 (16)	.35 (59)	-.13 (54)

Table 7

Squared and standardized parameter estimates from rater-specific effects model (see Figure 2)

	A	B	C	D	E	F	G	H	PC	GCC
effects common to both raters										
a (genetic)	.907	.959	.578	.941	.830	.785	.916	1	.987	.894
c (shared environment)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
e (non-shared environment)	.093	.040	.422	.058	.169	.215	.084	.000	.012	.105
rater-specific effects: parents										
a_p (genetic/rater contrast)	.112	.202	.534	.312	.620	.384	.331	.183	.268	.343
c_p (shared environment/rater bias)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
e_p (measurement error)	.270	.229	.187	.186	.258	.299	.385	.534	.271	.207
z_p (loading on common phenotype)	.616	.569	.279	.501	.125	.317	.304	.282	.335	.449
total genetic = $a_p + (z_p * a)$.671	.748	.695	.783	.724	.633	.609	.465	.599	.744
rater-specific effects: same teacher										
a_t (genetic/rater contrast)	.304	.000	.298	.189	.680	.542	.356	.286	.408	.470
c_t (shared environment/rater bias)	.000	.235	.100	.162	.000	.094	.129	.267	.247	.125
e_t (measurement error)	.105	.286	.401	.072	.193	.064	.183	.144	.077	.048
z_t total genetic = $a_t + (z_t * a)$.591	.478	.199	.577	.127	.298	.331	.303	.268	.356
total genetic = $a_t + (z_t * a)$.840	.458	.413	.732	.785	.776	.659	.589	.673	.788
significance of nested models										
drop all environment (c, e, c_p , c_t)	1.000	.663	.982	.717	.910	.982	.974	.609	.613	.951
drop common phenotype (a, c, e)	.000	.000	.000	.000	.040	.000	.000	.000	.000	.000
drop specific genetic (a_p , a_t)	.030	.022	.007	.003	.000	.000	.066	.195	.000	.000
drop a_t and c_t for different teachers	.401	.638	.247	.607	.013	.301	.951	.142	.010	.100

Figure 1

Boxplots showing distribution of GCC for parent (grey) and teacher (white) ratings in relation to language status at 6 years of age. The bold line denotes the median, the box shows middle 50% of sample, the fins denote 10th and 90th centiles, and circles correspond to outliers.

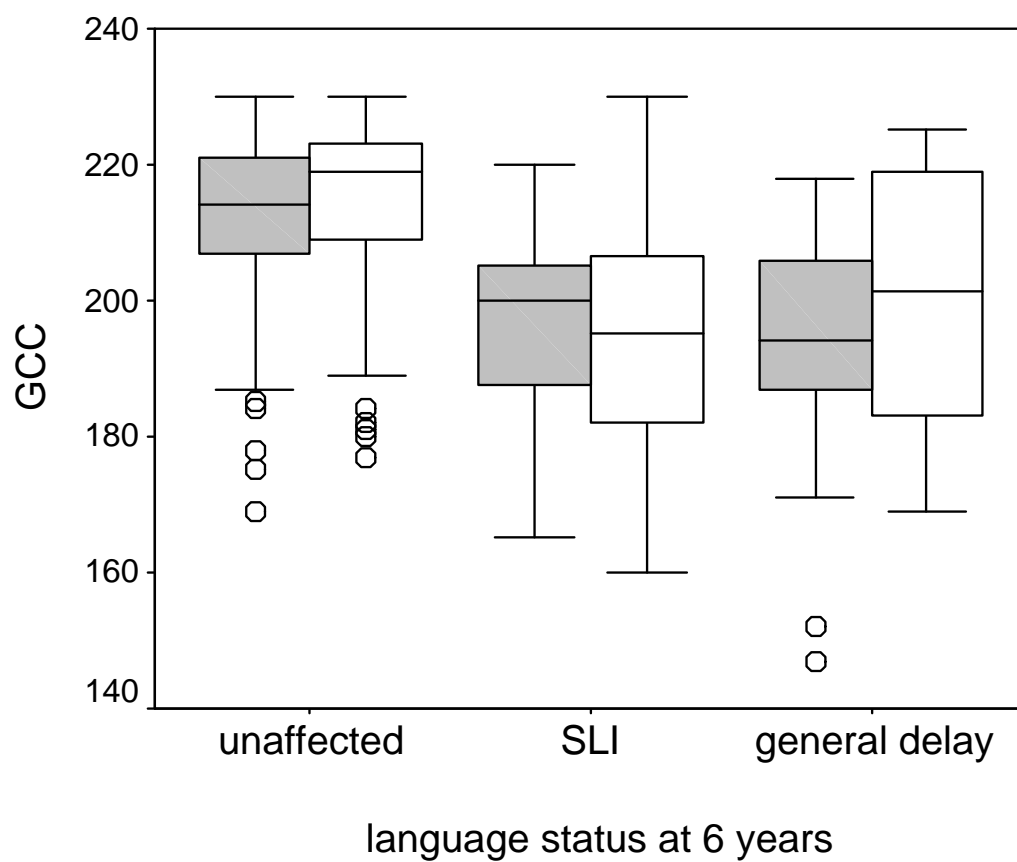


Figure 2

Rater-specific effects model (based on psychometric model of van der Valk et al., 2001)

